TikaAndNER

Named Entity Recognition (NER) with Tika

Named Entity Recognition is supported in *tika-parsers*, introduced in TIKA-1787. This page describes the steps required to configure and activate the NamedEntityParser.

Activate Named Entity Parser

Before moving ahead to configure NER implementations, *org.apache.tika.parser.ner.NamedEntityParser*, the parser responsible for handling the name recognition task needs to be enabled. This can be done with Tika Config XML file, as follows:

```
<?xml version="1.0" encoding="UTF-8"?>
<properties>
    <parsers>
        <parser class="org.apache.tika.parser.ner.NamedEntityParser">
            <mime>text/plain</mime>
            <mime>text/html</mime>
            <mime>application/xhtml+xml</mime>
        </parser>
    </parsers>
</properties>
```

This configuration has to be supplied in the later phases, so store it as 'tika-config.xml'.

Note: The NamedEntityParser parser does not restrict mimetypes, it uses Tika's auto detect parser to read text content from non-text streams.

Using Apache OpenNLP NER

The NE Parser is configured to use an implementation based on Apache OpenNLP. However, the NER models need to be added to the Tika's classpath to make this work.

The following table shows types of entities and the paths to place the model file.

Entity Type	Path for model	URL to get
PERSON	org/apache/tika/parser/ner/opennlp/ner-person.bin	http://opennlp.sourceforge.net/models-1.5/en-ner-person.bin
LOCATION	org/apache/tika/parser/ner/opennlp/ner-location.bin	http://opennlp.sourceforge.net/models-1.5/en-ner-location.bin
ORAGANIZATION	org/apache/tika/parser/ner/opennlp/ner-organization.bin	http://opennlp.sourceforge.net/models-1.5/en-ner-organization.bin
DATE	org/apache/tika/parser/ner/opennlp/ner-date.bin	http://opennlp.sourceforge.net/models-1.5/en-ner-date.bin
TIME	org/apache/tika/parser/ner/opennlp/ner-time.bin	http://opennlp.sourceforge.net/models-1.5/en-ner-time.bin
PERCENT	org/apache/tika/parser/ner/opennlp/ner-percentage.bin	http://opennlp.sourceforge.net/models-1.5/en-ner-percentage.bin
MONEY	org/apache/tika/parser/ner/opennlp/ner-money.bin	http://opennlp.sourceforge.net/models-1.5/en-ner-money.bin

Notes:

1. You can use any combination of the models. If you are interested in only the LOCATION names, then skip other NER models save LOCATION.
2. NER Models for other languages are also available http://opennlp.sourceforge.net/models-1.5/ . If you choose to use different language, use those URLs in the below script.

Tika App + OpenNLP NER in action

```
#Create a directory for keeping all the models.
#Choose any convenient path but make sure to use absolute path
export NER_RES=$HOME/tika/tika-ner-resources
mkdir -p $NER_RES
cd $NER_RES

PATH_PREFIX="$NER_RES/org/apache/tika/parser/ner/opennlp"
URL_PREFIX="http://opennlp.sourceforge.net/models-1.5"

mkdir -p $PATH_PREFIX

# using three entity types from the above table for demonstration
wget "$URL_PREFIX/en-ner-person.bin" -O $PATH_PREFIX/ner-person.bin
wget "$URL_PREFIX/en-ner-location.bin" -O $PATH_PREFIX/ner-location.bin
wget "$URL_PREFIX/en-ner-organization.bin" -O $PATH_PREFIX/ner-organization.bin

export TIKA_APP={your/path/to/tika-app}/target/tika-app-1.12-SNAPSHOT.jar

java -classpath $NER_RES:$TIKA_APP org.apache.tika.cli.TikaCLI --config=tika-config.xml -m http://people.apache.
org/committer-index.html

# Are there any metadata keys starting with "NER_" ?
```

Using Stanford CoreNLP NER

> The 'org.apache.tika.parser.ner.corenlp.CoreNLPNERecogniser' class provides runtime bindings to Stanford CoreNLP CRF classifiers for named entity recognition.

The following steps are necessary to use this NER implementation:

- Add Core NLP library and its dependencies to classpath
- Add models to class path
- Set NER Implementation to CoreNLP

NOTE: The latest release of Stanford CoreNLP requires JDK8.

Tika + CoreNLP in action

```
cd /$HOME/src
git clone https://github.com/thammegowda/tika-ner-corenlp.git
cd tika-ner-corenlp
mvn clean compile package assembly:single -PtikaAddon

#this should produce target/tika-ner-corenlp-addon-*-jar-with-dependencies.jar
export CORE_NLP_JAR=`find $PWD/target/tika-ner-corenlp-addon-*jar-with-dependencies.jar`

export TIKA_APP={your/path/to/tika-app}/target/tika-app-1.12-SNAPSHOT.jar

java -Dner.impl.class=org.apache.tika.parser.ner.corenlp.CoreNLPNERecogniser \
      -classpath $TIKA_APP:$CORE_NLP_JAR org.apache.tika.cli.TikaCLI \
      --config=tika-config.xml -m http://www.hawking.org.uk

# Observe metadata keys starting with NER_

# To use 3class NER model  (Default is 7 class model)

 java  -Dner.corenlp.model=edu/stanford/nlp/models/ner/english.all.3class.distsim.crf.ser.gz \
       -Dner.impl.class=org.apache.tika.parser.ner.corenlp.CoreNLPNERecogniser \
       -classpath $TIKA_APP:$CORE_NLP_JAR org.apache.tika.cli.TikaCLI \
       --config=tika-config.xml -m http://www.hawking.org.uk
```

The CoreNLP CRF classifier recognised the following from the text content of http://www.hawking.org.uk page:

```
NER_DATE: 2009
NER_DATE: 1963
NER_DATE: 1663
NER_DATE: 1982
NER_DATE: 1979
NER_LOCATION: Gonville
NER_LOCATION: Einstein
NER_LOCATION: London
NER_LOCATION: Cambridge
NER_LOCATION: Santa Cruz
NER_ORGANIZATION: Leiden University
NER_ORGANIZATION: NASA
NER_ORGANIZATION: CBE
NER_ORGANIZATION: Brief History of Time
NER_ORGANIZATION: University of California
NER_ORGANIZATION: Cambridge Lectures Publications Books Images Films
NER_ORGANIZATION: Caius College
NER_ORGANIZATION: Royal Society
NER_ORGANIZATION: About Stephen The Computer Stephen
NER_ORGANIZATION: US National Academy of Science
NER_ORGANIZATION: Department of Applied Mathematics
NER_ORGANIZATION: ESA
NER_ORGANIZATION: The Universe
NER_ORGANIZATION: Sally Tsui Wong-Avery Director of Research
NER_ORGANIZATION: the University of Cambridge
NER_ORGANIZATION: Theoretical Physics
NER_ORGANIZATION: Baby Universe
NER_PERSON: Einstein
NER_PERSON: P. Oesch
NER_PERSON: R. Bouwens
NER_PERSON: George
NER_PERSON: Stephen Hawking
NER_PERSON: Isaac Newton
NER_PERSON: D. Magee
NER_PERSON: Annie
NER_PERSON: G. Illingworth
NER_PERSON: Stephen
NER_PERSON: Dennis Stanton Avery
```

Using Regular Expressions

The **org.apache.tika.parser.ner.regex.RegexNERecogniser** provides an implementation based on Regular expressions. The following steps are required to use this implementation:

- Configure regular expressions in 'org/apache/tika/parser/ner/regex/ner-regex.txt'
- Set System property ner.impl.class to *org.apache.tika.parser.ner.regex.RegexNERecogniser*

Tika + RegexNER in action

```
# Create a regex file and add it to classpath
export NER_RES=$HOME/tika/tika-ner-resources
mkdir -p $NER_RES
cd $NER_RES
mkdir -p org/apache/tika/parser/ner/regex/

echo "PHONE_NUMBER=((\+\d{1,2}\s?)?\(?\d{3}\)?[\s.-]?\d{3}[\s.-]?\d{4})" > org/apache/tika/parser/ner/regex/ner-
regex.txt
echo "EMAIL=([a-zA-Z0-9.!#$%&'*+/=?^_`{|}~-]+@[a-zA-Z0-9](?:[a-zA-Z0-9-]{0,61}[a-zA-Z0-9])?(?:\.[a-zA-Z0-9](?:[a-
zA-Z0-9-]{0,61}[a-zA-Z0-9])?))" >> org/apache/tika/parser/ner/regex/ner-regex.txt

export TIKA_APP={your/path/to/tika-app}/target/tika-app-1.12-SNAPSHOT.jar

java -Dner.impl.class=org.apache.tika.parser.ner.regex.RegexNERecogniser \
    -classpath $NER_RES:$TIKA_APP org.apache.tika.cli.TikaCLI \
    --config=tika-config.xml -m  http://www.cs.usc.edu/faculty_staff/faculty

# Observe values of keys NER_PHONE_NUMBER and NER_EMAIL
```

Creating a custom NER

- Create a class and implement *org.apache.tika.parser.ner.NERecogniser*
- Set class name as value to system property *ner.impl.class* similar to Regex or CoreNLP

Chaining all the above at once

Multiple class names can be provided by setting the system property *ner.impl.class* to a comma separtes class names

Example : *-Dner.impl.class = org.apache.tika.parser.ner.opennlp.OpenNLPNERecogniser,org.apache.tika.parser.ner.regex.RegexNERecogniser*